Association for Information Systems
AIS Electronic Library (AISeL)

Research Papers | ECIS 2018 Proceedings

11-28-2018

HOW RESEARCH DATA MANAGEMENT CAN CONTRIBUTE TO EFFICIENT AND RELIABLE SCIENCE

Armel Lefebvre
Utrecht University, a.e.j.lefebvre@uu.nl

Elizabeth Schermerhorn
Utrecht University, e.a.schermerhorn@students.uu.nl

Marco Spruit
Utrecht University, m.r.spruit@uu.nl

Follow this and additional works at: https://aisel.aisnet.org/ecis2018_rp

Recommended Citation
Lefebvre, Armel; Schermerhorn, Elizabeth; and Spruit, Marco, "HOW RESEARCH DATA MANAGEMENT CAN CONTRIBUTE TO EFFICIENT AND RELIABLE SCIENCE" (2018). *Research Papers*. 35.
https://aisel.aisnet.org/ecis2018_rp/35

This material is brought to you by the ECIS 2018 Proceedings at AIS Electronic Library (AISeL). It has been accepted for inclusion in Research Papers by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

HOW RESEARCH DATA MANAGEMENT CAN CONTRIBUTE TO EFFICIENT AND RELIABLE SCIENCE

Research paper

Lefebvre, Armel, University of Utrecht, Utrecht, the Netherlands, a.e.j.lefebvre@uu.nl

Schermerhorn, Elizabeth, University of Utrecht, Utrecht, the Netherlands, e.a.schermerhorn@students.uu.nl

Spruit, Marco, University of Utrecht, Utrecht, the Netherlands, m.r.spruit@uu.nl

Abstract

Research data management (RDM) is an emergent discipline which is increasingly receiving attention from universities, funding agencies and academic publishers. While data management (DM) benefits from a large corpus of data governance and management frameworks adapted to industry, its academic counterpart RDM still struggles at identifying, organizing and implementing the main functions of RDM. In this study we explore the status of research data management at two research organizations in the Netherlands. We identify the main roles and tasks involved in research data governance, services and research. We show that, while the application of the DAMA-DMBOK functions and RDM structures are overlapping, RDM is coping with fundamentally different organizational structures and roles than the roles and tasks listed in professional DM frameworks. As RDM is developed to make science more efficient and reliable, it is questionable whether its current structure is effective. Based on interviews with data managers, researchers and librarians we identified several issues. For instance, at the moment, researchers are responsible for tasks that depend on DM expertise that they, generally, do not possess. At the same time, research data governance as currently implemented fails to capture the complexity of (professional) data management. Similarly, research data support is not well integrated with the wide diversity of research projects. If not addressed, these issues may impede any progress towards open, efficient and reliable science.

Keywords: Research Data management, Research infrastructure, Open Data, Data stewardship, Open Science, Research Data Governance

1 Introduction

Funding agencies are promoting an "as open as possible, as closed as necessary" (Directorate-general for Research and Innnovation, 2016, p. 1) principle for research data availability. It is a matter of guaranteeing trustworthy science and more efficient allocation of resources by, for instance, encouraging scientific data reuse by scientists, businesses and citizens (European Commission, 2016a). This position shared among major research funders in Europe are driving larger investments in (European) research infrastructures, the integration of data management plans with grant proposals and a broader adoption of open access as an option to publish scholarly output (Pryor, 2012; Wallis *et al.*, 2013; Ayris *et al.*, 2016; European Commission, 2016b). Overall, these research governance efforts are shaping future directions of data management (DM) for publicly funded research projects.

However, while substantial efforts have been made to deploy Open Science at a European level to make science more efficient and reliable (as explained further in section 2.1), the basic act of publishing research data still encounters significant resistance from academics (Borgman, 2012; Tsai *et al.*, 2016). Actually, the problem of "opening" data, which is seen as beneficial for verification and reuse of existing scientific material (Peng, 2011; Lefebvre, Spruit and Omta, 2015; Mannheimer, Sterman and Borda, 2016), is a striking example of one of the challenges universities or research institutes are facing when offering data management support to researchers. On the one hand universities are implementing such programs by means of data management plans (DMPs) to secure public funding in the coming years (Simms *et al.*, 2016). On the other hand, researchers are coping with numerous ways of producing and analyzing scientific data, which makes research data management costly, complex and diverse (Shoshni and Rotem, 2009).

In previous research on DM in academia, several empirical studies report on results of surveys which show the reluctance of researchers to make data open access and the lack of proper means for preserving data (Tenopir *et al.*, 2011; Fecher, Friesike and Hebing, 2015). Although this type of studies is informative to obtain more insights about the stakeholders and problematic aspects of RDM (e.g. storage issues, cumbersome data curation, poor rewards for publishing data…), they provide limited knowledge about the actual deployment of RDM programs in research organizations.

Therefore, this study investigates the deployment of existing research data governance and RDM programs in two research institutions (a university: *UNI_CASE* and research and healthcare organization: *HEALTH_CASE*) in the Netherlands using an *exploratory, interpretive* case study approach. This approach is chosen to collect experiences from data managers, librarians and researchers using qualitative, semi-structured interviews. We use the DAMA-DMBOK (Mosley *et al.*, 2010) as a reference framework which standardizes best data management practices in industry as a lecture grid for interpreting research data management activities, roles and infrastructure in academia.

In short, our approach aims at answering the following research question: **How can research data management contribute to efficient and reliable science?** This question implies to define the current state of research data management and the roles which are taking part in RDM programs. More, as we investigate how RDM can (positively) contribute to Open Science, several aspects impeding the deployment of research data management in research institutions are discussed.

By investigating the "interactions" between RDM governance, RDM services and researchers as they currently occur in research institutions, we aim at providing insights on challenges of managing research data in a way which is compliant to the requirements of Open Science. To achieve that, we first depict the current situation of research data management in two research organizations in the Netherlands. Next, we focus on a subset of management and governance activities related to the research data lifecycle: the lifecycle of research data from creation to publication and preservation. Finally, we discuss the most frequent issues identified by combining the data policy screening and the two case studies.

2 Theoretical background

In this section, we provide some background knowledge about data governance, data management and Open Science (OS). As explained earlier, research organizations in the Netherlands are seeking to implement new research data governance rules to comply with requests from external stakeholders (typically funding agencies and publishers) for making valuable datasets or software available.

2.1 Open science

Open science has, for the European commission, two goals in addition to openness, these are (European Commission, 2016a):

- **Reliable science** relates to the verification of published results. It encourages effective data quality checks and, in general, better research governance and scientific integrity for more credible and reproducible science.
- **Efficient science** focuses on resources needed to produce scientific knowledge. Efficient science seeks to reuse existing scientific material, thus limiting resource duplication. Additionally, it encourages the use of (web)standards, versioning of research artefacts and promotes connected tools and platforms. The European Open Science cloud declaration is an illustration of the commitment for federating research infrastructures in Europe (Ayris et al., 2016).

More, data management has also implications for reliable and efficient science with applications outside academic research such as *data science*. For instance, the EDISON project, which formalized knowledge and skills of data scientists (Manieri *et al.*, 2016), and the BDVA reference model (Zillner *et al.*, 2017) have data management as a core priority. The availability of scientific data to the industry and citizens in Europe place research data management and Open Science in direct connection to open innovation (Chesbrough, 2012). To achieve this, two building blocks of RDM must be considered to maximize the quality of available research data:

Research Data Lifecycles describe the steps research data undergoes before, during and after the project is completed. There is no unified view on research data lifecycles (RDL) discussed in the literature. For instance, RDLs can be oriented towards data management or on data curation. They can also be vary from field to field, e.g. geology or social sciences (Higgins, 2008; Pryor, 2012). Although there is no unique reference RDL, several common steps can be extracted. These are (1) planning, (2) creation/collection, (3) processing, (4) analysis, (5) publication, (6) archival/dismissal, (7) reuse. The point of view of data curation is that research data might be repurposed for other projects. In that case, published data serves as input for another study, which justifies the use of a cycle.

Lately, **Data management plans (DMPs)** are submitted by researchers as part of a funding agreement with a public funding agency. A DMP outlines RDM-oriented activities during the whole lifecycle of the data i.e. from research design to archival (Corti *et al.*, 2014; European Commission, 2016b; Simms *et al.*, 2016)**.** Until now, funders are not assessing DMPs as part of a grant proposal but, some of them, like NWO, make DMPs a prerequisite for the actual subsidy of granted projects.

2.2 DAMA-DMBOK: Data governance and management

The DAMA-DMBOK is an industry standard for data management created by DAMA international (Otto, 2011). It arranges data management into functions to which corresponds groups of activities: planning, control, development and operational activities (Mosley *et al.*, 2010, p. 24). For instance, Data governance (DG) is "an organizational approach to data and information management that formalizes a set of data policies and procedures to encompass the full life cycle of data, from acquisition to use and to disposal" (Korhonen *et al.*, 2013, p. 11). The policies and procedures are applicable on strategic, tactical and operational decision-making levels in an organization (Mosley *et al.*, 2010; Korhonen *et al.*, 2013). DG is the core process of Data Management (DM). DG coordinates nine other DM functions by exercising planning and control activities (Mosley *et al.*, 2010). The functions coor-

dinated by DAMA-DMBOK represents key areas for managing data, these are: (1) data architecture management, (2) data development, (3) database operations management, (4) data security management, (5) reference and master data management, (6) data warehousing and business intelligence management, (7) document and content management, (8) meta-data management, (9) data quality management. The *DAMA-DMBOK* also states that IT professionals and business people (named data stewards) are involved in data management programs, each participating in one or more functions.

Next, DM is defined as "the planning, execution and oversight of policies, practices and projects that acquire, control, protect, deliver, and enhance the value of data and information assets" (Mosley *et al.*, 2010). It is to note that the ambitions of research data management (RDM) are in-line with the definition of DM as given by the DAMA-DMBOK. In the end, RDM aims at enhancing the value of research data, despite the current struggles to identify quality or relevance metrics which could be applicable to datasets (Belter, 2014). RDM principles known as the **F**indable, **A**ccessible, **I**nteroperable and **R**eusable (FAIR) data (Wilkinson *et al.*, 2016) are guiding RDM infrastructure development and practices to generate reliable, quality research data. The FAIR principles are endorsed by major European and Dutch funders (European Commission, 2016b; NWO, 2017).

Accordingly, the initial case study design started with the selection of three DM functions from DAMA-DMBOK which relates to the main topics found in RDM policies. Then, we performed a matching based on RACI charts to identify relevant activities from RDM and link them to functions from the DAMA-DMBOK framework (see Table 1). More details about the matching are given in section 3.1. Below we present the three RDM functions which resulted from the screening of Dutch RDM policies.

Research data governance: noticeably, RDM governance policies inspected during this case study are not articulated around functions like DAMA-DMBOK. Instead, very practical points such as licensing, informed consents or funder requirements are addressed. In a glance, these activities should be dealt with by researchers (or higher-level Faculty management), who are responsible for their execution. So, in contrast to DAMA-DMBOK, no official equivalent to enterprise-council or data stewardship coordination are found, except in one data policy which is attributing (partial) auditing and policy development responsibilities to a "Research data office", a center of expertise.

RDM Functions	DAMA-DMBOK Definitions	RDM Activities
Governance **DAMA-DMBOK: Data governance**	"The exercise of authority and control (planning, monitoring, and enforcement) over the management of data assets" (Mosley *et al.*, 2010, p. 37)	Audit, Monitoring, Ethics, Legislation, Funder requirements, Establishing Policies and Procedures, Selection of Standards, Licensing, Informed Consent, Authorizations
Services **DAMA-DMBOK: Data operations activities**	"Planning, control and support for structured data assets across the data lifecycle, from creation and acquisition to archival and purge" (Mosley *et al.*, 2010, p. 130)	Raise awareness, Training, Guidelines, IT support, Data Management Plan support, Ownership support, Data stewardship support, Knowledge network
Researchers **DAMA-DMBOK: Data development**	"Designing, implementing and maintaining solutions to meet the data needs of the enterprise" (Mosley *et al.*, 2010, p. 88)	Collection, Creation, Processing, Analysis, Publication, Archival, Reuse, Retention, Documentation, Quality, Storage, Back-up, Versioning

Table 1. A preliminary matching between RDM activities identified from policy screening and formal definitions of data governance, data operations and data development functions found in the DAMA-DMBOK framework.

Research data management services regroup (1) IT support, which can be close to the researcher or centralized, and (2) centralized library services giving RDM workshops for researchers, support for writing DMPs etc. These activities correspond to *data operations*, a function of the DAMA-DMBOK which concentrates on data handling during the entire data lifecycle.

Researchers are matched to the *data development* function as their activities resemble the most to the development activities of the *data development function* of DAMA-DMBOK. Although this function

is the closest to what researchers do with data, as shown by the software and databases published by researchers in many domains, it is also the less satisfying matching to the DMBOK of all three categories. It indicates that the closest match for this category has no formal function defined in the DMBOK. Existing DMBOK functions do not cover well the activities of researchers in a single function. For instance, data development is not well suited to researchers simply using software or platforms. More, it has an emphasis on data modelling, for instance, which is not a typical activity done by researchers.

3 Research Design

We collected evidence from two different type of data sources, as using multiple sources is recommended by Benbasat et al. (1987) for exploratory case studies where cross-case analysis is performed. Here, cross-case analysis was the option we chose for the reasons stated earlier, i.e. collecting evidence from a diverse organization and from a more centralized one. First, we analyzed research data management policies to identify the roles, tasks and responsibilities that frequently occur in RDM. Next, we conducted 22 interview sessions in one university and its university medical center.

Figure 1 shows the relations between the main concepts exposed earlier. Research data governance is developing policies for data management planning and assigns tasks to researchers and data management services. RDM services' tasks are mainly related to support. Researchers retain most of the responsibilities of managing research data. RDM is structured as such to help the organization satisfy goals of Open Science. These are Reliability and Efficiency of science. These two goals are implemented in the Netherlands in research organizations (such as UNI_CASE and HEALTH_CASE).

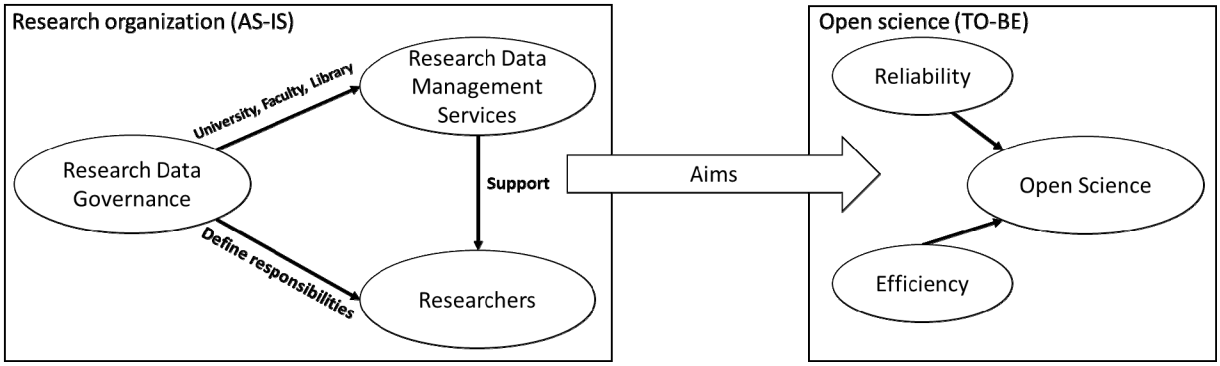

Figure 1. Initial RDM functions retained for devising an interview protocol.

3.1 Governance: Policy review

To identify several functional areas of the DMBOK which might be of interest for exploring RDM, Dutch RDM policies were screened prior to the case studies. The sample of policies consisted of documents from all universities in the Netherlands, including *UNI_CASE,* and *HEALTH_CASE*. The policy review captures the roles and tasks of data management in research institutions as validated by the administration board of each organization.

In total, 13 Dutch RDM policies were scanned using a RACI chart, a type of *Responsibility Assignment Matrix* (Wende and Otto, 2007). As explained by Wende & Otto (2007), RACI stands for **R**esponsible, **A**ccountable, **C**onsulted and **I**nformed. The matrix's rows are tasks related to data management and the columns are the roles. The values (R/A/C or I) indicate which type of responsibility a given role received for a task. A summary of the analysis is shown in section 4.1.

- Responsible is unique per row in the matrix as it refers to the person who performs the task

- Accountable refers to the role which has ultimate decision power over a task
- Consulted is a role providing input before completion of the task, consult is an optional role
- Informed is simply a role notified of the completion of a task, is optional

The advantage of RACI chart for policy review is that their interpretation is straightforward. The comparison of policies from different organizations make inconsistencies in the attribution of responsibility very clear. A disadvantage of this technique, as we experienced, is that the matching between matrices containing unstandardized or equivocal roles and tasks is a cumbersome process for which the correctness is hard to evaluate.

3.2 Exploratory case studies

Next, we collected qualitative data by means of interviews with practitioners during two exploratory, interpretive, case studies (Benbasat, Goldstein and Mead, 1987; Klein and Myers, 1999). The first case study (*UNI_CASE)* is a research organization in the Netherlands. The second case is a healthcare and research organization (*HEALTH_CASE*) closely bound to *UNI_CASE* but having a patient care mission missing from *UNI_CASE*. A subset of the interviewees (total n=23) were recruited during a network meeting, attended by research data managers, held at *UNI_CASE in April 2017*. This networking event attracted participants from *RDM Services* and *RDM Research*. We had no interviews with members of the executive board of the university (or faculties) who are formally responsible for developing policies and authorizing their application at an organization-wide level. Nevertheless, people involved in the creation of governance frameworks were interviewed in both organizations.

3.3 Sites selection

There are two main strategies guiding the selection of case study sites. The first one is opportunistic. We had access to a network of data managers working in both organizations which facilitated the recruitment of interviewees. The second strategy was theoretically grounded. UNI_CASE and HEALTH_CASE differ mainly due the sensitivity of the data analyzed for research purposes. HEALTH_CASE started a RDM program earlier than UNI_CASE for that reason and it has a stronger centralization of RDM solutions than UNI_CASE, which is a university with very diverse research faculties. They are not strictly speaking two independent organizations. *HEALTH_CASE* is the *Faculty of medical sciences* of the university *UNI_CASE*. The separation between the faculty of medical sciences and the other faculties is common to all universities in the Netherlands. So, the two aspects of RDM in context: diversity for UNI_CASE and, centralization and sensitivity for HEALTH_CASE

3.4 Interview protocol

An initial group of interviewees was contacted during a data management network meeting at UNI_CASE. The other interviewees were recommended at the end of the first interviews for their expertise in one or more RDM activities. We designed the interview protocol based on the retained DAMA-DMBOK functions (data governance, development and operations), the data stewardship lifecycle (RDL) and tasks found in data governance policies of universities and medical centers. Also, we retained two aspects of data management to discuss with the participants. The first aspect relates to the three DAMA-DMBOK functions. The interviewees present how data is governed and supported in their organization/division. At the start of each interview, participants were invited to describe their job (or role) to facilitate the matching of their profile to two of the three categories retained for this study (i.e. service or research). The second aspect is focused on the connection between their tasks and typical stages of the research data lifecycle. Each interviewee had to discuss some of the tasks they are performing at each stage of the RDL presented to them as a printed picture.

3.5 Data collection

We conducted the interviews between May 2017 and June 2017. In total, we held 22 interview sessions. The interviews were recorded and transcribed for further analysis in *NVivo 11.4,* a qualitative data analysis software. 21 sessions took place in Dutch, one interview in English. Table 2 shows the role and domain of each participant. The roles refer to the job titles of the participants.

UN	Role	Domain	RDM	HE	Role	Domain	RDM
U1	1 Consultant	Library services	Services	**H1**	1 Researcher	Primary Care	Research
U2	1 Information Manager	Faculty Level IT	Services	**H2**	1 Product Owner 1 IT Architect	Central IT	Services
U3	1 Coordinator	Psychology	Services	**H3**	2 Data managers	Brain division	Services
U4	1 ICT Manager	Epidemiology	Services	**H4**	1 Researcher	Neuropsychiatry	Research
U5	1 Librarian	Library services	Services	**H5**	1 Manager/ Researcher	Infrastructure	Services/ Research
U6	1 Librarian	Library services	Services	**H6**	1 Researcher	Child division	Research
U7	1 ICT Manager	Epidemiology	Services	**H7**	2 Data managers	Vital functions	Services
U8	1 ICT Manager	Central IT	Services	**H8**	1 Data manager	Brain division	Services
U9	1 Information Manager	Geosciences	Services	**H9**	1 Researcher	Cell screening	Research
U10	1 Researcher	Finance	Research				
U11	1 Consultant	Library services	Services				
U12	1 Researcher	Epidemiology	Research				

Table 2 Background of interviewees and mapping to RDM main function

4 Data Analysis and Interpretation

In this section we describe the two types of data analyses that were executed. First, the policy screening (see section 4.1) explains how policy documents where mapped to RACI charts and how a meta-analysis of 8 Dutch RDM policies was made. Then details about the interviews are given in section 4.2).

4.1 Research Data Policy screening

The goal of the data policy screening is to collect roles and tasks as they are implemented in several research institutions. The present analysis gives a limited overview of the content of RDM policies. In Table 3, a value (R or A, as C and I hardly appear) is added only when roles and their task(s) are mentioned in the documents. Also, a minimal frequency of roles and tasks is used to filter out roles/tasks that differ too much. They had to appear at least in 2 documents to be added to this table. Some of the roles that are not passing the threshold are *data protection officer, data steward,* meaning that only one policy (out of 8) formalized responsibilities about at least one of these two roles.

The small number of Dutch policies we screened has several reasons. First, the absence of data governance policy at 2 Dutch research organizations. In addition, three other documents were guidelines. not formal policies attributing tasks to defined roles. These three documents could not be analyzed with RACI charts, as only explicit role assignments were considered for screening. Finally, other medical centers in the Netherlands have a *research code of conduct* but no RDM policy could be retrieved online.

4.2 Case studies

The interviews were recorded and transcribed into text files for further analysis in *NVivo*. With *N*Vivo, the interviews were classified per site (UNI_CASE and HEALTH_CASE). A first coding lead to a mapping of statements to predefined categories (e.g. data lifecycle, reuse, general remarks etc.) using 75 codes for UU_CASE and 67 nodes for HEALTH_CASE. Next, a more open coding process, as defined by Heath & Cowley (2004), was conducted to search for more fine-grained information in the interviews. At a statement level, 355 nodes referred to quotes from interviewees from both organizations. The statement level coding was used to annotate keywords (e.g. agreement, awareness, data quality) and context (e.g. no formal training, data management costs etc.). In the next sections, we substantiate the different tasks with experiences from data managers, researchers and librarians. All these categories are derived from the research data policy screening. They are presented in the order they appear in Table 3. We start with the creation of data management plan (in section 4.3).

Task\Role	Function	Researcher	P.I.	Faculty	Executive Board	Library
Create Data Management Plan	*Research*	*R (5)*	*R (2)*	-	-	-
Handle Data Lifecycle (Stewardship)	*Research*	*R (5), A (1)*	-	-	-	-
Support IT Infrastructure	*Service*	-	-	*R (2)*	*R (2)*	-
Support Training and Data Handling	*Service*	-	-	*R (1)*	*R (1)*	*R (1)*
Monitor and Audit	*Governance*	-	-	*R (2)*	*R (2)*	-
Develop and Implement policy	*Governance*	-	-	*R (2)*	*R (2)*	-

Table 3. The distributions of roles found per tasks, based on 8 RACI charts. The number between () indicates how many times this relation has been encountered among the 8 policies analyzed (status of February 2017).

4.3 Create Data Management Plan

Researchers are responsible for creating a data management plan (DMP). Results from the policy screening and interviewees indicate that this rule applies also at UNI_CASE. The separation between RDM services and researchers appears clearly for data management planning. At UNI_CASE, RDM services are available for help but researchers are still fully responsible for writing it. As explained by an RDM consultant: "We offer a review service. We do not write the DMP for the researcher except if it is a big project, then they can hire us. We are hired on a temporary basis to make the DMP, but it is not the goal that everybody does so…" **[U1]**. Another consultant adds: "We provide a DMP check but they [researchers] need to fill it in themselves. We do not have any mandate to validate something. They remain responsible for their DMP and its good quality" **[U11].**

Located closer to researchers, an information manager explained that support for DMP is rerouted to the library: "Now we have to submit a DMP with grant proposals. In that case they [researchers] know where to find us and we know where to find the library" **[U9]**. Moreover, none of the interviewees from *UNI_CASE* could remember a DMP that would have been written without the request of a funder though *UNI_CASE* has officially enforced the use of DMPs since January 2016 for all new scientific projects.

In HEALTH_CASE, however, the involvement of data managers in RDM planning seems stronger. Two data managers explained: "We support researchers and we are involved in the creation of the DMP, […]. We help with the data collection tools, the software with which data is collected. This is done in collaboration with the researchers, for surveys and the data from electronic health records" **[H3]**. A manager of a computing facility for bioinformaticians confirms the closer involvement with researchers in their organization: "Fortunately, researchers come more often to us. But the problems

arrive, through the principal investigator (PI), to the IT department if it is about storage or something. Fortunately, we have a direct line with IT, so we know where to redirect them to an information point where they know more about it. We called the person who is now busy with audits and data managements plans a data steward, this person is our information point now" **[H5].**

4.4 Handle Data Lifecycle

A DMP describes how data will be handled during and after a research project. Handling research data is a responsibility of researchers (or their principal investigators). Again, some major differences occur in the implementation of this rule in the two organizations.

First, data collection and data creation are separated activities. **Data collection** refers to data gathered from already existing data sources and **data creation** occurs when researchers use data which did not exist before. **Data collection** is frequent at *HEALTH_CASE* where operational data from the medical center serves as research data for researchers. The connection between patient data and research data is done by data managers and a centralized research data platform (i.e. a data warehouse). Data managers perform quality checks on the data as it is filled in by multiple people (e.g. nurses, doctors) and are not meant for research but for care in the first place. Then, upon request from a researcher, an anonymized subset of the data is transferred. Two data managers state that the fact that operational medical data is used for research purposes have implications on data governance: "Data collection is really important and often different than a university because in a university you start with a study, you select participants and they all consent to the study and they participate. In a medical center, there is already a lot of data that you want to use for your research. This raises other questions regarding governance…" **[H7]**. **[H5]** and **[H6]** add that data can also be created by a *wet lab person* manipulating instruments. After, this data can be integrated with data from health records stored in the *research data platform*.

At *UNI_CASE,* **[U3]** sometimes helps researchers with extracting data from databases and performs some data cleaning. In other cases, data might be collected directly via services from financial companies which have agreements with the university **[U10]**. Another scenario is that data is generated by external providers, as it is the case for **[U7]**, a data manager for a longitudinal study with a large cohort: "there are always things that are unclear, and which are not known by the agency [which manages surveys], so we ask them to put memo's or codes if they don't know. And then it starts with the retrieval of memos and codes, we have developed all kind of rules around that because the study started a long time ago. […] we have also a small data cleaning part". On the sides of librarians **[U5]** and **[U11]** are involved in data collection when it concerns reuse of published datasets exclusively.

Next, as two data managers from the brain division state: "Data processing, analysis and publication are done by the researchers themselves" **[H3].** This statement summarizes the roles involved in **processing and analysis** tasks. Nevertheless, in the case of consortia, tasks are distributed across organizations and processing might be done at another organization, as explained by one researcher in Neuropsychiatry **[H4]**. If the analysis is done by third parties, researchers might have some difficulties to understand how the data was analyzed: "If we did a part of the processing or data analysis for them [researchers], they come back to us asking what [analyses] did you do again?" **[H5].** In **[U12]**'s team (epidemiology), data managers are processing and documenting the data and their role are perceived as really important.

Following the analysis step, **archival** was an issue at UNI_CASE and HEALTH_CASE. There are diverging opinions about what should be done. At HEALTH_CASE, there are rules for preserving raw data for 5 years and data for verification for 15 years. The difference in retention time is due to the high costs a 15-year preservation of raw material would entail **[H5]**. An additional hindrance is the absence of strict rules or guidelines for archival **[H7]** and researchers must decide on what they archive and what not. At UNI_CASE, data managers and researchers are experiencing many similar troubles. The general rule in the Netherlands are that underlying data must be preserved for 10 years after publication. For **[U1], [U8]** and **[U11],** nobody else than researchers are solely responsible for deciding about what data to dismiss after a project ended.

Finally, **reuse and publication** suffer from similar issue in both organizations. The Faculty of Geosciences of UNI_CASE, for instance, shares and reuses data created by others as the analysis methods in their field rely on measurements from different locations on earth **[U9]**. Then, archival techniques impede reusability, not all raw data can be stored, and the archive consists of a processed bulk **[U2]**. According to **[H4]**, even internal reuse is challenging and requires contacting different people in their department to obtain information about where to find relevant datasets. None of the interviewees has put data open at time of the interviews. In the case of long-term, longitudinal studies, the cause was ancient informed consents: "In the case of [longitudinal] cohorts, it does not happen. The informed consents signed 20 years ago stated that the data would not be made openly available. So, we are not authorized to do so. Sometimes we must explain that to journals too, they want us to make the data open access, until now we were successful. It is not possible. The data is available upon request to individual researchers." [**U12].** At *HEALTH*_CASE, a data manager affirms: "I do not know how it works with open data and patient information […] I do not know what the rules are, therefore we are not doing it" [**H7].** In short, the archival issues, privacy, longitudinal studies with no informed consents for open data, limited space in repositories, commercial agreements, unclear regulations and reluctance from researchers are the causes listed by the interviewees to not (be able to) make data open by default.

4.5 Support IT Infrastructure, Training and Data Handling

Two initiatives from the *central IT departments* exist to deploy new IT infrastructures for research. In HEALTH_CASE, it is mainly a data warehouse aggregating information from different internal sources. It was developed to provide researchers with integrated, anonymized operational data. Researchers have to fill-in a request form where they describe their research questions before receiving the information they need from data managers. As such, data managers at HEALTH_CASE are there to protect sensitive information hosted in the medical center.

In UNI_CASE, technologies vary per projects, disciplines and faculties. The *Central IT department* is developing a storage environment in collaboration with **[U3]**'s team. At the same time, **[U2]** explained that storage is provided as a paid service for departments inside the faculty but that they are not (yet) competitive against external storage solutions. Meanwhile, **[U11]** stated that library services do not offer any storage too, they only focus on support. As a consultant explained: "RDM support is a combination of a university-wide program for IT and research, the central IT department, the T&T (teaching and research) department, legal affairs, privacy and other departments. It is not a formal organization but a real collaboration inside UNI_CASE" **[U1]**. From the perception of researchers, this scaffold around RDM appears far away from their more urgent issues. One researcher from HEALTH_CASE: "I had no contact with them [RDM support]. Who should be the contact point… right now everything is really centralized but they are not the people that can help me if I have questions about protocols or data management plans. They are too far away and too generic […]" **[H1].**

Nonetheless, training for managing data is perceived as a benefit for researchers: "I do think that at the start of the career you should receive a workshop or some education on how to properly do this, this is important. A lot of mistakes have been made at the beginning which I wouldn't do now…" [**U11]**. But there are no DM training or workshops specialized per discipline of type of analyses in both organizations. The existing workshops are said to be generic and are directed at raising awareness of researchers about RDM topics, they lack some depth as argued by **[H1]** and **[H9]**.

4.6 Monitoring and Developing policies and Guidelines

Monitoring and audit of research data occurred in one case: when patient data is involved. There is no active monitoring or auditing on how research data management is deployed in the two organizations by the roles that are responsible for these tasks (Faculty board and Executive boards). **[U11]** said it will take time before monitoring is going to happen, the reason given is that the implementation of RDM goes slowly and more time is needed to put these control mechanisms in place.

5 Results

In this section we comment on the results of the policy screening and the interviews. Furthermore, some limitations of our approach are addressed in section 5.3. We also relate our findings to the IS roadmap for open data implementation (Link *et al.*, 2017) to further guide the development of the roadmap.

5.1 RDM policies

What appears from the screening outcomes, summarized in Table 3, are that these documents do not clearly state any accountability, consulted or informed roles (apart from one policy in which researchers are accountable for data handling). Hence, it is not clear how tasks are divided among RDM stakeholders. They differ per project and are not directly linkable to roles present in the DAMA-DMBOK, an industry standard for data management.

Besides, there are several layers of data policies that are to be developed: deans and executive boards are both responsible for developing data management. They differ in scope though. While the executive board develops a central (university-wide) policy, refinements are delegated to faculties which are assumed to be able to define clearer and more explicit regulations covering characteristics of their own research data. It is to note that there is, to a certain extent, a discrepancy between the tasks related to researchers and those involving executive boards (faculties and university). It can be seen from Table 3 that there is an agreement to attribute DMP and data handling tasks to researchers (5 out of 8) but few data policies explicitly mentioned infrastructure support, training and further policy developments for which libraries or executive boards from the university or faculties are responsible.

Further, Table 3 shows that there are no specific RDM roles defined. Unlike the DAMA-DMBOK which provides a set of 32 roles relating to DM responsibilities at enterprise and business unit levels, RDM reshuffles well-known academic positions (e.g. researchers, principal investigator (PI), deans), and appears to not regulate additional roles such as *data manager* or *data steward*, even if these roles actively collaborate with researchers and interact with research data during its lifecycle.

Finally, no monitoring and audit procedures are present, while *data management* (as defined in section 2.2) has a major monitoring component according to industry standards. The monitoring task, as shown in Table 3, refers to *responsibilities* to monitor policy development and data handling, not on formal procedures or metrics achieving this. From the case studies, we can also confirm that there is no active monitoring in place, except for strict legal reasons (i.e. privacy): when patient data is involved.

5.2 Research data at the Medical Centre and the University

We have seen that, for both organizations, a division of RDM in three functions: governance, services and research, suffices to categorize most of the roles involved. But it is also noticeable that RDM services and RDM research can be further refined. For RDM services, two profiles emerged. The first, the *governance supporter,* profile supports researchers from an *Open Science* perspective. Their activities are constrained to *post* analysis data curation, data management planning support and raising awareness about open data. Another profile, the *research supporter*, belonging to RDM services is closer to operational activities that researchers are conducting. The former profile is related to *governance* support more than support for researchers. The latter coincides more with what can be expected from *research* support and encompasses *data managers* and *data stewards*.

The two goals of Open Science, efficiency and reliability, are not the most prominent drivers for *research supporters* at RDM services in *HEALTH_CASE* and *UNI_CASE*. As said earlier, data security is the primary interest for the *research data platform* at HEALTH_CASE and the federated storage of UNI_CASE serves archival and encryption needs of a longitudinal study. The same can be said about *data management plans*. Data management planning is done when required by funders, and the rule of UNI_CASE to enforce DMP for each new project had no concrete effect. This might indicate two

things: researchers do not see the benefit of data management planning (if no funding depends on it) and central data policies from UNI_CASE have a weak impact on changing the behavior of researchers regarding data planning.

Governance is structured top-down in both organizations. UNI_CASE initiated a central policy which needs to be refined by each faculty, which is still ongoing work at this time. However, there is no evidence that faculties are the most optimal decision layer when it comes to managing research data. For instance, the institute where [U7] is located has no contact with the faculty as their presence in that faculty is purely administrative, the type of research and data differs from the rest of the faculty. Other interviewees agreed that the type of research (e.g. quantitative, qualitative) and type of data (e.g. commercial, medical, experimental, simulated) are significantly more impacting the services needed to plan and handle this data accordingly. Hence, faculty boards might not be a suitable basis for further elaborating on responsibilities and tasks as those tend to share more commonalities with equivalent analysis and data than other departments belonging to the same faculty or institute.

5.3 Discussion and Limitations

There are several limitations to this study. First the data collection is limited in scope. Indeed, the preliminary RDM division in functions and roles (see Table 3) is established based on a limited number of data policies and two case studies in the Netherlands. Identifying to which extent similar RDM implementations exist in research institutions requires further research. Nevertheless, other countries where RDM has gained some maturity appear to use a similar division between functions. In the UK, 79% of the institutional policies "mentioned and specified" a role for RDM support at an institutional level, but only 37% of the 57 policies define clear control (i.e. review) and responsibilities (Horton and DCC, 2016). Although control is an activity group of data governance according to DAMA-DMBOK, it seems that it is not systematically enforced in policies from research institutions, neither in the UK or in the Netherlands.

When the scope is broadened to North American research institutions, a study from Tenopir *et al.*, (2014) highlighted that library services are indeed not intervening at a technical level during data analysis and provide (or are planning to provide) support for curation and data management plans. These activities are similar to our findings, where RDM services offered by librarians is mostly restrained to consultancy. We can add that consultancy is insufficiently regulated by RDM Governance which tends to assign most of the responsibilities to researchers or academic management staff without clear rules for opening data. This absence of guidelines has an impact on research supporters as well. This makes research supporters (e.g. data stewards, data managers) unsure about how opening data should be done, legally and technically. These findings lead us to contextualize the issues of opening research data into a broader agenda. At a higher-level, Ponte (2015) indicated that technical quality and sustainability of the open data are issues threatening the open data ecosystem. While the research institution we investigated clearly work on increasing data quality (open or not), it is questionable whether the whole enterprise is sustainable if tasks and responsibilities remain vague or if control is not exercised thoroughly.

In addition to data quality and sustainability issues, infrastructure and privacy were two other aspects which are perceived as being complex matters by the interviewees. Actually, the motivations for open data implementation introduced by Link et al. (2017) echo to the goals of Open Science and RDM functions presented earlier. This allows us to further examine the open data implementation agenda under the light of our findings. The authors classified open data implementation into two dimensions: motivation and implementation. The four motivations we elaborate on here are "mandated sharing" , "benefits to the research process" and "extending the life of research data" (Link *et al.*, 2017, p. 563). We do not discuss "career impact" as the focus of our case study results is more in-line with the three other motivations:

- Mandated sharing: Funders, publishers or universities encourage open data.

- Benefits to the research process: reliability of science by facilitating the reproduction of results.
- Extending the life of research data refers to the notion of efficiency of science introduced earlier.

Motivation		**Mandated Sharing**	**Research Process** ***(Reliability)***	**Life of Research Data** **(*Efficiency)***
Implementation	Governance	No DMP filled in if not requested by funders	Institutional policies to be overridden by faculties and departments, no standardization of custom policies	Questions arise with data collected from companies (copyrights) or with (old) informed consents (privacy), no clear accountability.
	Socio-technical System	Centralized solutions are developed when data is considered sensitive, open data itself is not sufficient	Institutional storage not competitive with cloud storage, but cloud storage is not used if data is sensitive	No clear rules for archiving research data
	Standards	No standard way of handling data during the lifecycle	Operational data or outsourced data collection might not be standardized	Some "standards" are tailor-made for a project, they are not cross-discipline
	Data Quality	Data might be collected from different sources where researchers have few control on its quality	Freedom when it comes to archive data, researchers are solely responsible	Internal reuse difficult, not immediate insights how data was generated
	Ethics	Privacy is a reason not to share	Lack of guidelines or infrastructure for opening sensitive data	There are cases were operational medical data is used by researchers which influence the possibility to publish it

Table 4. Classification of our findings according to the IS roadmap for open data (Link et al., 2017, p. 595).

Table 4 shows several issues identified during our case studies and how they relate to the roadmap of open data implementation (Link *et al.*, 2017). This way, concrete practical issues which are in-line with the IS roadmap are given to further advance the discussions. These are situations that are to be expected and which negatively impact the to-be situation of Open Science. These findings form discussion points to nurture IS research to become a major role player in open data implementation and RDM to contribute to Open Science.

6 Conclusion and Further research

This paper investigated how research data management can contribute to Open Science. Open science has two goals: reliability and efficiency of science. Both goals rely on RDM to be attained. At the same time, RDM is lagging industry standards on several aspects, mainly data governance which stays vague on data management planning and control. For that reason, the open by default strategy is not applied due to regulatory and operational issues that, if properly addressed, could ease the data publication process. **Efficiency of science** through better regulated RDM can be achieved by involving *research supporters* and make their responsibilities clearer in data policies, in which they are currently not well represented. **Reliable science** is challenged by other operational issues such as data archival and privacy. *Governance supporters* are more perceived, on the researchers' side, as Open Science champions without proper infrastructure to make RDM work. Their responsibilities and tasks must be formalized, at least as *consulted* or *informed* roles to foster data publication.

Further research should collect more evidence from other research institutions worldwide following the policy screening and exploratory case study approach. More roles, tasks and functions can be discovered and refine the three main RDM functions found in the two organizations. Eventually, research data management can benefit from guidelines and assessment instruments grounded in evidence obtained from larger scale policy screenings and the of RDM in diverse research institutions.

References

Ayris, P. *et al.* (2016) *Realising the European Open Science Cloud*.

Belter, C. W. (2014) 'Measuring the value of research data: A citation analysis of oceanographic data sets', *PLoS ONE*, 9(3). doi: 10.1371/journal.pone.0092590.

Benbasat, I., Goldstein, D. K. and Mead, M. (1987) 'The Case Research Strategy in Studies of Information Systems', *MIS Quarterly*, 11(3), p. 369. doi: 10.2307/248684.

Borgman, C. L. (2012) 'The conundrum of sharing research data', *Journal of the American Society for Information Science and Technology*, pp. 1059–1078. doi: 10.1002/asi.22634.

Chesbrough, H. (2012) 'Open Innovation: Where We've Been and Where We're Going', *Research-Technology Management*, 55(4), pp. 20–27. doi: 10.5437/08956308X5504085.

Corti, L. *et al.* (2014) *Managing and sharing research data: A guide to good practice*.

Directorate-general for Research and Innnovation (2016) *Open Research Data as the default: Frequently Asked Questions about the exte nsion of the Open Research Data Pilot*.

European Commission (2016a) *EU Open Innovation, Open Science, Open to the World*, *European Comission*. doi: 10.2777/061652.

European Commission (2016b) *Guidelines on Data Management in Horizon 2020*. Available at: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf (Accessed: 1 March 2016).

Fecher, B., Friesike, S. and Hebing, M. (2015) 'What drives academic data sharing?', *PLoS ONE*. Edited by R. S. Phillips. Public Library of Science, 10(2), p. e0118053. doi: 10.1371/journal.pone.0118053.

Heath, H. and Cowley, S. (2004) 'Developing a grounded theory approach: a comparison of Glaser and Strauss', *International Journal of Nursing Studies*, 41, pp. 141–150. doi: 10.1016/S0020-7489(03)00113-5.

Higgins, S. (2008) 'DCC Curation Lifecycle Model', *International Journal of Digital Curation*, 3(1), pp. 134–140. doi: 10.2218/ijdc.v2i2.30.

Horton, L. and DCC (2016) *Overview of UK Institution RDM Policies', Version 6 August 2016, Digital Curation Centre*. Available at: http://www.dcc.ac.uk/resources/policy-and-legal/institutional-data-policies.

Klein, H. H. H. K. and Myers, M. D. (1999) 'A set of principles for conducting and evaluating interpretive field studies in information systems', *MIS quarterly*, 23(1), pp. 67–93. doi: 10.2307/249410.

Korhonen, J. J. *et al.* (2013) 'Designing Data Governance Structure : An Organizational Perspective', *Journal on Computing*, 2(4), pp. 11–17. doi: 10.5176/2251-3043.

Lefebvre, A., Spruit, M. and Omta, W. (2015) 'Towards reusability of computational experiments capturing and sharing research objects from knowledge discovery processes', in *IC3K 2015 - Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*.

Link, G. *et al.* (2017) 'Contemporary Issues of Open Data in Information Systems Research: Considerations and Recommendations', *Communications of the Association for Information Systems*.

Manieri, A. *et al.* (2016) 'Data science professional uncovered: How the EDISON project will contribute to a widely accepted profile for data scientists', in *Proceedings - IEEE 7th International Conference on Cloud Computing Technology and Science, CloudCom 2015*, pp. 588–593. doi: 10.1109/CloudCom.2015.57.

Mannheimer, S., Sterman, L. B. and Borda, S. (2016) 'Discovery and Reuse of Open Datasets: An Exploratory Study', 5, p. e1091. doi: 10.7191/jeslib.2016.1091.

Mosley, M. *et al.* (2010) *DAMA guide to the data management body of knowledge*. Technics Publications.

NWO (2017) *Data management protocol*. Available at: https://www.nwo.nl/en/policies/open+science/data+management (Accessed: 25 November 2017).

Otto, B. (2011) *A MORPHOLOGY OF THE ORGANISATION OF DATA GOVERNANCE*, *ECIS 2011 Proceedings*. doi: 10.1007/978-3-8348-9953-8.

Peng, R. D. (2011) 'Reproducible research in computational science.', *Science*. NIH Public Access, 334(6060), pp. 1226–7. doi: 10.1126/science.1213847.

Ponte, D. (2015) 'Enabling an Open Data Ecosystem', *ECIS 2015 Research-in-Progress Papers*.

Pryor, G. (2012) *Managing research data*. Facet Publishing.

Shoshni, A. and Rotem, D. (2009) *Scientific Data Management: Challenges, Technology, and Deployment*.

Simms, S. *et al.* (2016) 'The Future of Data Management Planning : Tools , Policies , and Players', *International Digital Curation Conference (IDCC16)*, (February 22-25), p. 10.

Tenopir, C. *et al.* (2011) 'Data sharing by scientists: practices and perceptions.', *PloS one*, 6(6), p. e0118053. doi: 10.1371/journal.pone.0021101.

Tenopir, C. *et al.* (2014) 'Research data management services in academic research libraries and perceptions of librarians', *Library & Information Science Research*. JAI, 36(2), pp. 84–90. doi: 10.1016/J.LISR.2013.11.003.

Tsai, A. C. *et al.* (2016) 'Promises and pitfalls of data sharing in qualitative research', *Social Science & Medicine*. doi: 10.1016/j.socscimed.2016.08.004.

Wallis, J. C. *et al.* (2013) 'If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology', *PLoS ONE*. Edited by L. A. Nunes Amaral. Public Library of Science, 8(7), p. e67332. doi: 10.1371/journal.pone.0067332.

Wende, K. and Otto, B. (2007) 'A contingency approach to data governance', *Proceedings, 12th International Conference on Information Quality (ICIQ-07), Cambridge, USA*, p. 14. doi: 10.1002/nml.241.

Wilkinson, M. D. *et al.* (2016) 'The FAIR Guiding Principles for scientific data management and stewardship', *Scientific Data*. Nature Publishing Group, 3, p. 160018. doi: 10.1038/sdata.2016.18.

Zillner, S. *et al.* (2017) *European Big Data Value Strategic Research and Innovation Agenda*.